

The Nonconvex Geometry of Low-Rank Matrix Optimizations with General Objective Functions

Qiuwei Li and Gongguo Tang*

Department of Electrical Engineering and Computer Science,
Colorado School of Mines, Golden, CO 80401

November 10, 2016

Abstract

This work considers the minimization of a general convex function $f(X)$ over the cone of positive semi-definite matrices whose optimal solution X^* is of low-rank. Standard first-order convex solvers require performing an eigenvalue decomposition in each iteration, severely limiting their scalability. A natural nonconvex reformulation of the problem factors the variable X into the product of a rectangular matrix with fewer columns and its transpose. For a special class of matrix sensing and completion problems with quadratic objective functions, local search algorithms applied to the factored problem have been shown to be much more efficient and, in spite of being nonconvex, to converge to the global optimum. The purpose of this work is to extend this line of study to general convex objective functions $f(X)$ and investigate the geometry of the resulting factored formulations. Specifically, we prove that when $f(X)$ satisfies restricted strong convexity and smoothness, each critical point of the factored problem either corresponds to the optimal solution X^* or is a strict saddle point where the Hessian matrix has a negative eigenvalue. Such a geometric structure of the factored formulation ensures that many local search algorithms can converge to the global optimum with random initializations.

1 Introduction

Consider a general semi-definite program (SDP) where a convex objective function $f(X)$ is minimized over the cone of positive semi-definite (PSD) matrices:

$$\underset{X \in \mathbb{R}^{n \times n}}{\text{minimize}} f(X) \quad \text{subject to} \quad X \succeq 0. \quad (1.1)$$

For this problem, even fast first-order methods such as the projected gradient descent algorithm [18, 25] require performing an expensive eigenvalue decomposition in each iteration. These expensive operations form the major computational bottleneck of the algorithms and prevent them from scaling to scenarios with millions of variables, a typical situation in a diverse of applications, including phase retrieval [20, 64], quantum state tomography [1, 33], user preferences prediction [29, 53, 57, 65], and pairwise distances estimation in sensor localization [13, 14].

When the SDP (1.1) admits a low-rank solution X^* , in their pioneer work [19], Burer and Monteiro proposed to factorize the variable $X = UU^T$, where $U \in \mathbb{R}^{n \times r}$ with $r \ll n$, and solved a factored nonconvex problem

$$\underset{U \in \mathbb{R}^{n \times r}}{\text{minimize}} g(U) := f(UU^T). \quad (1.2)$$

*Q. Li and G. Tang were supported by the NSF Grant CCF-1464205.

For standard SDPs with a linear objective function and several linear constraints, they also argued that when the factorization $X = UU^T$ is overparameterized, *i.e.*, $r > r^* := \text{rank}(X^*)$, any local minimum of (1.2) corresponds to the solution X^* , provided some regularity conditions are satisfied. Unfortunately, these regularity conditions are generally hard to verify for specific SDPs arising in applications. Recent work [16] removed these regularity conditions and showed that the factored objective function $g(U)$ almost never has any spurious local optima for general linear objective functions. Our work differs in that the convex objective function $f(X)$ is generally not linear and there are no additional linear constraints. In addition to showing the nonexistence of spurious local minima, we also demonstrate that critical points that do not correspond to global optima are strict saddle points, ensuring the global convergence of simple gradient descent algorithms.

A special class of optimizations that admit low-rank solutions stem from regularizing matrix inverse problems using the trace or nuclear norm [52], which have found numerous applications in machine learning [34], signal processing [17], and control [45]. The statistical performance of such optimizations in recovering a low-rank matrix have been studied extensively in literature using convex analysis techniques [23]. For example, it has information-theoretically optimal sampling complexity [24], achieves minimax denoising rate [21] and satisfies tight oracle inequalities [22]. In spite of its optimal performance, trace norm regularization cannot be scaled to solve the practical problems that originally motivate its development even with specialized first-order algorithms. This was realized since the advent of this field and low-rank factorization method was proposed as an alternative to convex solvers [19]. When coupled with stochastic gradient descent, low-rank factorization leads to state-of-the-art performance in practical matrix completion problems [32, 68].

The past two years have seen renewed interest in the Burer-Monterio factorization for solving trace norm regularized inverse problems. With technical innovations in analyzing the nonconvex landscape of the factored objective function, several recent works have shown that with exact parameterization (*i.e.*, $r = r^*$) the factored objective function $g(U)$ in trace norm regularized matrix inverse problems has no spurious local minima or degenerate saddle points [11, 12, 32, 39, 62]. An important implication is that local search algorithms such as gradient descent and its variants are able to converge to the global optimum with even random initialization [12].

We generalize this line of work by assuming a general objective function $f(X)$ in the optimization (1.1), not necessarily coming from a matrix inverse problem. The generality allows us to view the factored problem (1.2) as a way to solve the convex optimization (1.1) to global optimality, rather than a new modeling method. This perspective, also taken by Burer and Monterio in their original work, frees us from rederiving the statistical performances of the factored optimization (1.2). Instead, its performance inherits from that of the convex optimization (1.1), whose performance can be developed using a suite of powerful convex analysis techniques accumulated from several decades of research. As a specific example, the optimal sampling complexity and minimax denoising rate of trace norm regularization need not to be rederived once one knows the equivalence between the convex and the factored formulations. In addition, our general analysis technique also sheds light on the connection between the geometries of the convex program (1.1) and its nonconvex counterpart (1.2) as discussed in Section 3.

Our governing assumption on the objective function $f(X)$ is $2r$ -restricted m -strong convexity and M -smoothness. More precisely, the Hessian of $f(X)$ satisfies

$$m\mathbf{I} \preceq \nabla^2 f(X) \preceq M\mathbf{I} \quad (1.3)$$

for some positive numbers M and m and any PSD matrix X with $\text{rank}(X) \leq 2r$. Here \mathbf{I} is an identity matrix of appropriate sizes. This assumption is standard in matrix inverse problem [4, 47]. We show that under this assumption, each critical point of the factored objective function $g(U)$ either corresponds to the low-rank global solution of the original convex program (1.1) or is a strict saddle point where the Hessian $\nabla^2 g(U)$ has a strictly negative eigenvalue. These results are summarized in the following theorem:

Theorem 1. *Suppose the function $f(X)$ in (1.1) is twice continuously differentiable and satisfies $2r$ -restricted m -strong convexity and M -smoothness condition (1.3) with positive numbers M and m satisfying*

$$\frac{M}{m} \leq 1.15. \quad (1.4)$$

Assume X^* is an optimal solution of the minimization (1.1) with $\text{rank}(X^*) = r^*$. Set $r \geq r^*$ in (1.2). Let U be any critical point of $g(U)$ satisfying $\nabla g(U) = \mathbf{0}$. Then U either corresponds to a square-root factor of X^* , i.e.,

$$X^* = UU^T; \quad (1.5)$$

or is a strict saddle point of the factored problem (1.2). More precisely, let $U^* \in \mathbb{R}^{n \times r}$ such that $X^* = U^*U^{*T}$. and set $D = U - U^*R$ with $R = \text{argmin}_{R: RR^T = \mathbf{I}_r} \|U - U^*R\|_F^2$, then the curvature of $\nabla^2 g(U)$ along D is strictly negative:

$$[\nabla^2 g(U)](D, D) < -0.074m(\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2 \|D\|_F^2 \quad (1.6)$$

with $r' := \text{rank}(U)$ and $r' \vee r^* := \max\{r', r^*\}$. This further implies

$$\lambda_{\min}(\nabla^2 g(U)) < -0.074m(\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2. \quad (1.7)$$

Several remarks follow. First, the matrix D is the direction from the saddle point U to its closest global optimal factor U^*R of the same size as U . Second, we can simplify the expression $(\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2$ in (1.7) depending on the relative values of r' and r^* :

$$(\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2 = \begin{cases} \sigma_{r^*}(U^*)^2 & \text{if } r' < r^*; \\ (\sigma_{r'}(U) + \sigma_{r^*}(U^*))^2 & \text{if } r' = r^*; \\ \sigma_{r'}(U)^2 & \text{if } r' > r^*. \end{cases}$$

For all these cases, $(\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2$ is strictly positive, implying that U is a strict saddle point. Note that our result covers both over-parameterization where $r > r^*$ and exact parameterization where $r = r^*$. Third, we can recover the rank- r^* global minimizer X^* of (1.1) by running local search algorithms on the factored function $g(U)$ if we know an upper bound on the rank r^* . The strict saddle property ensures that many iterative algorithms, for example, stochastic gradient descent [31], trust-region method [58, 60], and gradient descent with sufficiently small stepsize [41], all converge to a square-root factor of X^* , even with random initialization. Last but not least, our main result only relies on the restricted strong convexity and smoothness property. Therefore, in addition to low-rank matrix recovery problems [67, 22, 37] and phase retrieval [60, 55, 20, 50], it is also applicable to many other low-rank matrix optimization problems with non-quadratic objective functions, including 1-bit matrix completion also known as the logistic PCA [28, 40], robust PCA with complex noise [48, 66], Poisson PCA [54], and other low-rank models with generalized loss functions [63]. For SDPs with additional linear constraints, as those studied in [19, 16], one can combine the original objective function with a least-squares term that penalizes the deviation from the linear constraints. As long as the penalization parameter is large enough, the solution is equivalent to that of the standard SDP and hence is also covered by our main theorem.

We end this section by introducing some notations used throughout the paper. Denote $[n]$ as the collection of all positive integers up to n . The symbols \mathbf{I} and $\mathbf{0}$ are reserved for the identity matrix and zero matrix/vector, respectively. A subscript is used to indicate its size when this is not clear from context. We call a matrix PSD, denoted by $X \succeq 0$, if all its eigenvalues are nonnegative. The notation $X \succeq Y$ means $X - Y \succeq 0$, i.e., $X - Y$ is PSD. The set of $r \times r$ orthogonal matrices is denoted by $\mathbb{O}_r = \{R \in \mathbb{R}^{r \times r} : RR^T = \mathbf{I}_r\}$. Matrix norms such as the spectral, nuclear, and Frobenius norms are denoted respectively by $\|\cdot\|$, $\|\cdot\|_*$ and $\|\cdot\|_F$.

The gradient of a scalar function $f(Z)$ with a matrix variable $Z \in \mathbb{R}^{m \times n}$ is an $m \times n$ matrix, whose (i, j) th entry is $[\nabla f(Z)]_{i,j} = \frac{\partial f(Z)}{\partial Z_{ij}}$ for $i \in [m]$, $j \in [n]$. Alternatively, we can view the gradient as a linear form $[\nabla f(Z)](G) = \langle \nabla f(Z), G \rangle = \sum_{i,j} \frac{\partial f(Z)}{\partial Z_{ij}} G_{ij}$ for any $G \in \mathbb{R}^{m \times n}$. The Hessian of $f(Z)$ can be viewed as a 4th order tensor of size $m \times n \times m \times n$, whose (i, j, k, l) th entry is $[\nabla^2 f(Z)]_{i,j,k,l} = \frac{\partial^2 f(Z)}{\partial Z_{ij} \partial Z_{kl}}$ for $i, k \in [m]$, $j, l \in [n]$. Similar to the linear form representation of the gradient, we can view the Hessian as a bilinear form defined via $[\nabla^2 f(Z)](G, H) = \sum_{i,j,k,l} \frac{\partial^2 f(Z)}{\partial Z_{ij} \partial Z_{kl}} G_{ij} H_{kl}$ for any $G, H \in \mathbb{R}^{m \times n}$. Yet another way to represent

the Hessian is as an $mn \times mn$ matrix $[\nabla^2 f(Z)]_{i,j} = \frac{\partial^2 f(Z)}{\partial x_i \partial x_j}$ for $i, j \in [mn]$, where x_i is the i th entry of the vectorization of Z . We will use these representations interchangeably whenever the specific form can be inferred from context. For example, in the strong convexity and smoothness condition (1.3), the Hessian is apparently viewed as an $n^2 \times n^2$ matrix and the identity \mathbf{I} is of size $n^2 \times n^2$.

For a matrix-valued function $\phi : \mathbb{R}^{p \times q} \rightarrow \mathbb{R}^{m \times n}$, it is notationally easier to represent its gradient (or Jacobian) and Hessian as multi-linear operators. For example, the gradient, as a linear operator from $\mathbb{R}^{p \times q}$ to $\mathbb{R}^{m \times n}$, is defined via $[\nabla[\phi(U)](G)]_{ij} = \sum_{k \in [p], l \in [q]} \frac{\partial[\phi(U)]_{ij}}{\partial U_{kl}} G_{kl}$ for $i \in [m], j \in [n]$ and $G \in \mathbb{R}^{p \times q}$; the Hessian, as a bilinear operator from $\mathbb{R}^{p \times q} \times \mathbb{R}^{p \times q}$ to $\mathbb{R}^{m \times n}$, is defined via $[\nabla^2[\phi(U)](G, H)]_{ij} = \sum_{k_1, k_2 \in [p], l_1, l_2 \in [q]} \frac{\partial^2[\phi(U)]_{ij}}{\partial U_{k_1 l_1} \partial U_{k_2 l_2}} G_{k_1 l_1} H_{k_2 l_2}$ for $i \in [m], j \in [n]$ and $G, H \in \mathbb{R}^{p \times q}$. Using this notation, the Hessian of the scalar function $f(Z)$ of the previous paragraph, which is also the gradient of $\nabla f(Z) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$, can be viewed as a linear operator from $\mathbb{R}^{m \times m}$ to $\mathbb{R}^{m \times n}$ denoted by $[\nabla^2 f(Z)](G)$ and satisfies $\langle [\nabla^2 f(Z)](G), H \rangle = [\nabla^2 f(Z)](G, H)$ for $G, H \in \mathbb{R}^{m \times n}$.

2 Problem Formulations and Preliminaries

This paper considers the problem (1.1) of minimizing a convex function $f(X)$ over the PSD cone. Let X^* be an optimal solution of (1.1) of rank r^* . When the PSD variable X is reparameterized as

$$X = \phi(U) := UU^T,$$

where $U \in \mathbb{R}^{n \times r}$ with $r \geq r^*$ is a rectangular, matrix square root of X , the convex program is transformed into the factored problem (1.2) whose objective function is $g(U) := f(\phi(U))$. Inspired by the lifting technique in constructing SDP relaxations, we refer to the variable X as the lifted variable, and the variable U as the factored variable. Similar naming conventions apply to the optimization problems, their domains, and objective functions.

The nonlinear parametrization $\phi(U)$ makes $g(U)$ a nonconvex function and introduces additional critical points (*i.e.*, those U with $\nabla g(U) = 0$ that are not global optima of the factored optimization (1.2)). Our goal is to show that the critical points either correspond to X^* or are strict saddle points where the Hessian has a strictly negative eigenvalue.

2.1 Metrics in the Lifted and Factored Spaces

Since for any U , $\phi(U) = \phi(UR)$ where $R \in \mathbb{O}_r$, the domain of the factored objective function $g(U)$ is stratified into equivalent classes and can be viewed as a quotient manifold [2]. The matrices in each of these equivalent classless differ by an orthogonal transformation (not necessarily unique when the rank of U is less than r). One implication is that, when working in the factored space, we should consider all factorizations of X^* :

$$\mathcal{A}^* = \{U^* \in \mathbb{R}^{n \times r} : \phi(U^*) = X^*\}.$$

A second implication is that when considering the distance between two points U_1 and U_2 , one should use the distance between their corresponding equivalent classes:

$$d(U_1, U_2) := \min_{R_1 \in \mathbb{O}_r, R_2 \in \mathbb{O}_r} \|U_1 R_1 - U_2 R_2\|_F = \min_{R \in \mathbb{O}_r} \|U_1 - U_2 R\|_F, \quad (2.1)$$

where the second equality follows from the rotation invariance of $\|\cdot\|_F$. Under this notation, $d(U, U^*) = \min_{R \in \mathbb{O}_r} \|U - U^* R\|_F$ represents the distance between the class containing a critical point $U \in \mathbb{R}^{n \times r}$ and the optimal factor class \mathcal{A}^* . The second minimization problem in the definition (2.1) is known as the orthogonal Procrustes problem, whose optimal R is characterized by the following lemma:

Lemma 1. [35] *An optimal solution for the orthogonal Procrustes problem:*

$$R = \operatorname{argmin}_{\tilde{R} \in \mathbb{O}_r} \|U_1 - U_2 \tilde{R}\|_F^2 = \operatorname{argmax}_{\tilde{R} \in \mathbb{O}_r} \langle U_1, U_2 \tilde{R} \rangle$$

is given by $R = LP^T$, where the orthogonal matrices $L, P \in \mathbb{R}^{r \times r}$ are defined via the singular value decomposition of $U_2^T U_1 = L \Sigma P^T$. Moreover, we have $U_1^T U_2 R = (U_2 R)^T U_1 \succeq 0$ and $\langle U_1, U_2 R \rangle = \|U_1^T U_2\|_*$.

For any two matrices $U_1, U_2 \in \mathbb{R}^{n \times r}$, the following lemma proved in Section A relates the distance $\|\phi(U_1) - \phi(U_2)\|_F$ in the lifted space to the distance $d(U_1, U_2)$ in the factored space.

Lemma 2. Assume that $U_1, U_2 \in \mathbb{R}^{n \times r}$ has ranks r_1, r_2 , respectively. Then

$$\|\phi(U_1) - \phi(U_2)\|_F \geq (\sigma_{r_1 \vee r_2}(U_1) + \sigma_{r_1 \vee r_2}(U_2))d(U_1, U_2),$$

where $\sigma_\ell(\cdot)$ denotes the ℓ -th largest singular value and $r_1 \vee r_2 = \max\{r_1, r_2\}$.

2.2 Consequences of Restricted Strong Convexity

The parameterization $\phi(U) = UU^T$ introduces nonconvexity into the factored problem (1.2). We are interested in studying how it transforms the landscape of the lifted objective function $f(X)$. We make the assumption that the landscape of $f(X)$ in the lifted space is bowl-shaped, at least along matrices of rank at most $2r$, as indicated by the $2r$ -restricted strong convexity and smoothness assumption (1.3). An immediate consequence of this assumption is that if (1.1) has an optimal solution X^* with $\text{rank}(X^*) \leq r$, then there is no other optimum of (1.1) with rank less than or equal to r :

Proposition 1. Suppose the function $f(X)$ is twice continuously differentiable and satisfies $2r$ -restricted m -strongly convexity condition in (1.3). Assume X^* is an optimum of the minimization (1.1) with $\text{rank}(X^*) \leq r$. Then X^* is the unique global optimum of (1.1) of rank at most r .

Proof. We prove it by contradiction. Suppose there exists another optimum X of (1.1) with $\text{rank}(X) \leq r$ and $X \neq X^*$. Then the second order Taylor's expansion reads

$$f(X) = f(X^*) + \langle \nabla f(X^*), X - X^* \rangle + \frac{1}{2}[\nabla^2 f(\tilde{X})](X - X^*, X - X^*),$$

where $\tilde{X} = tX^* + (1-t)X$ for some $t \in [0, 1]$ and $[\nabla^2 f(\tilde{X})](X - X^*, X - X^*)$ evaluates the Hessian bilinear form along the direction $X - X^*$. The KKT conditions for the convex optimization (1.1) states that $\nabla f(X^*) \succeq 0$ and $\nabla f(X^*)X^* = \mathbf{0}$, implying that the second term in the above Taylor expansion $\langle \nabla f(X^*), X - X^* \rangle = \langle \nabla f(X^*), X \rangle \geq 0$ since X is feasible. Further, since $\tilde{X} = tX^* + (1-t)X, t \in [0, 1]$ is PSD and has $\text{rank}(\tilde{X}) \leq 2r$, by the $2r$ -restricted m -strongly convexity assumption (1.3), we have

$$[\nabla^2 f(\tilde{X})](X - X^*, X - X^*) \geq m\|X - X^*\|_F^2.$$

Combining all, we get

$$f(X) \geq f(X^*) + \frac{1}{2}m\|X - X^*\|_F^2 > f(X),$$

since $X - X^* \neq \mathbf{0}$, which is a contradiction. \square

The restricted strong convexity and smoothness assumption (1.3) reduces to the Restricted Isometry Property (RIP) when the objective function is quadratic. This is apparent from the following equivalent form of the assumption:

$$\left\| \frac{2}{M+m} \nabla^2 f(X) - \mathbf{I} \right\| \leq \frac{M-m}{M+m} \quad (2.2)$$

that holds for any PSD matrix X of rank at most $2r$. This further implies a restricted orthogonality property:

$$\left| \left[\frac{2}{M+m} \nabla^2 f(X) \right] (G, H) - \langle G, H \rangle \right| \leq \frac{M-m}{M+m} \|G\|_F \|H\|_F \quad (\widehat{\text{RIP}})$$

that again holds for any PSD X of rank at most $2r$. Similar to the standard RIP, the $\widehat{\text{RIP}}$ also claims that the operator $\frac{2}{M+m} \nabla^2 f(X)$, when evaluated on a restricted set of low-rank matrices, preserves geometric structures.

3 Transforming the Convex Landscape

Our primary interest is to understand how the landscape of the lifted objective function $f(X)$ is transformed by the factored parameterization $\phi(U) = UU^T$, particularly how its global optimum is mapped to the factored space, how other types of critical points are introduced, and what are their properties. As a constrained convex optimization, all critical points of (1.1) are global optima and are characterized by the necessary and sufficient KKT condition [18]:

$$\nabla f(X^*) \succeq \mathbf{0}, \nabla f(X^*)X^* = \mathbf{0}, X^* \succeq \mathbf{0}. \quad (3.1)$$

Proposition 1 further shows that, as a consequence of the $2r$ -restricted strong convexity, such global optimum is unique among all PSD matrices of rank at most r . The factored optimization (1.2) is unconstrained, whose critical points are specified by the zero gradient condition:

$$\nabla g(U) = \nabla f(\phi(U))U = \mathbf{0}. \quad (3.2)$$

To classify the critical points, we compute the Hessian bilinear form $[\nabla^2 g(U)](D, D)$ as:

$$[\nabla^2 g(U)](D, D) = 2\langle \nabla f(\phi(U)), DD^T \rangle + 4[\nabla^2 f(\phi(U))](DU^T, DU^T). \quad (3.3)$$

Roughly speaking, the Hessian quadratic form has two terms – the first term involves the gradient of $f(X)$ and the Hessian of $\phi(U)$, while the second term involves the Hessian of $f(X)$ and the gradient of $\phi(U)$. Since $\phi(U + D) = \phi(U) + UD^T + DU^T + DD^T$, the gradient of ϕ is the linear operator $[\nabla \phi(U)](D) = UD^T + DU^T$ and the Hessian bilinear operator applies as $[\nabla^2 \phi(U)](D, D) = DD^T$. Note in (3.3) the second quadratic form is always nonnegative since $\nabla^2 f(\cdot) \succeq 0$ due to the convexity of $f(X)$.

For any critical point U of $g(U)$, the corresponding lifted variable $\phi(U) = UU^T$ is PSD and satisfies $\nabla f(\phi(U))\phi(U) = 0$. On one hand, if $\phi(U)$ further satisfies $\nabla f(\phi(U)) \succeq 0$, then in view of the KKT conditions (3.1) and noting $\text{rank}(\phi(U)) \leq r$, we must have $\phi(U) = X^*$, the global optimum of (1.1). On the other hand, if $\phi(U) \neq X^*$, implying $\nabla f(\phi(U)) \not\succeq 0$ due to the necessity of (3.1), then additional critical points can be introduced into the factored space. Fortunately, $\nabla f(\phi(U)) \not\succeq 0$ also implies that the first quadratic form in (3.3) might be negative for a properly chosen direction D . To sum up, the critical points of $g(U)$ can be classified into two categories: the global optima in the optimal factor set \mathcal{A}^* with $\nabla f(\phi(U)) \succeq 0$ and those with $\nabla f(\phi(U)) \not\succeq 0$. For the latter case, by choosing a proper direction D , we will argue that the Hessian quadratic form (3.3) has a strictly negative eigenvalue, and hence moving along D in a short distance will decrease the value of $g(U)$, implying that they are strict saddle points and are not local minima.

We argue that a good choice of D is the direction from current U to its closest point in the optimal factor set \mathcal{A}^* . Formally, $D = U - U^*R$ where $R = \arg\min_{\tilde{R} \in \mathbb{O}_r} \|U - U^*\tilde{R}\|_F$ is the optimal rotation for the orthogonal Procrustes problem. Plugging D into the first term of (3.3), we simplify it as

$$\begin{aligned} \langle \nabla f(\phi(U)), DD^T \rangle &= \langle \nabla f(\phi(U)), U^*U^{*T} - U^*RU^T - U(U^*R)^T + UU^T \rangle \\ &= \langle \nabla f(\phi(U)), U^*U^{*T} \rangle \\ &= \langle \nabla f(\phi(U)), U^*U^{*T} - UU^T \rangle, \end{aligned} \quad (3.4)$$

where in the second inequality the last three terms involving U were canceled and in the last equality the term $-UU^T$ was reintroduced both due to the critical point property $\nabla f(\phi(U))U = 0$. To build intuition on why (3.4) is negative while the second term in (3.3) remains small, we consider a simple example: the matrix Principal Component Analysis (PCA) problem.

Example 1. Matrix PCA Problem. Consider the PCA problem for symmetric PSD matrices:

$$\underset{X \in \mathbb{R}^{n \times n}}{\text{minimize}} f_{\text{PCA}}(X) := \frac{1}{2} \|X - X^*\|_F^2 \quad \text{subject to } X \succeq 0, \quad (3.5)$$

where X^* is a symmetric PSD matrix of rank r^* . Apparently, the optimal solution is $X = X^*$. Now consider the factored problem:

$$\underset{U \in \mathbb{R}^{n \times r}}{\text{minimize}} g(U) := f_{\text{PCA}}(UU^T) = \frac{1}{2} \|UU^T - U^*U^{*T}\|_F^2,$$

where $U^* \in \mathbb{R}^{n \times r^*}$ satisfies $\phi(U^*) = X^*$. Our goal is to show that any critical point U such that $\phi(U) \neq X^*$ is a strict saddle point. Since $\nabla f_{\text{PCA}}(X) = X - X^*$, by (3.4), the first term of $[\nabla^2 g(U)](D, D)$ in (3.3) becomes

$$\begin{aligned} 2\langle \nabla f_{\text{PCA}}(\phi(U)), DD^T \rangle &= 2\langle \nabla f_{\text{PCA}}(UU^T), U^*U^{*T} - UU^T \rangle \\ &= 2\langle UU^T - U^*U^{*T}, U^*U^{*T} - UU^T \rangle = -2\|UU^T - U^*U^{*T}\|_F^2 \end{aligned} \quad (3.6)$$

is strictly negative.

The second term of $[\nabla^2 g(U)](D, D)$ is $4\|DU^T\|_F^2$ since $\nabla^2 f_{\text{PCA}}(X) = \mathbf{I}$. We next argue that $DU^T = \mathbf{0}$. For this purpose, let $X^* = Q \text{diag}(\boldsymbol{\lambda}) Q^T = \sum_{i=1}^{r^*} \lambda_i \mathbf{q}_i \mathbf{q}_i^T$ be the eigenvalue decomposition of X^* , where $Q = [\mathbf{q}_1 \ \cdots \ \mathbf{q}_{r^*}] \in \mathbb{R}^{n \times r^*}$ has orthonormal columns and $\boldsymbol{\lambda} \in \mathbb{R}^{r^*}$ is composed of positive entries. Similarly, let $\phi(U) = V \text{diag}(\boldsymbol{\mu}) V^T = \sum_{i=1}^{r'} \mu_i \mathbf{v}_i \mathbf{v}_i^T$ be the eigenvalue decomposition of $\phi(U)$, where $r' = \text{rank}(U)$. The critical point U satisfies $-\nabla g(U) = 2(X^* - \phi(U))U = \mathbf{0}$, implying that

$$0 = (X^* - \sum_{i=1}^{r'} \mu_i \mathbf{v}_i \mathbf{v}_i^T) \mathbf{v}_j = X^* \mathbf{v}_j - \mu_j \mathbf{v}_j, j = 1, \dots, r'.$$

This means (μ_j, \mathbf{v}_j) forms an eigenvalue-eigenvector pair of X^* for each $j = 1, \dots, r'$. Consequently, $\mu_j = \lambda_{i_j}$ and $\mathbf{v}_j = \mathbf{q}_{i_j}$ and we can write $\phi(U) = \sum_{j=1}^{r'} \lambda_{i_j} \mathbf{q}_{i_j} \mathbf{q}_{i_j}^T = \sum_{j=1}^{r^*} \lambda_j s_j \mathbf{q}_j \mathbf{q}_j^T$. Here s_j is equal to either 0 or 1 indicating which of the eigenvalue-eigenvector pair $(\lambda_j, \mathbf{q}_j)$ appears in the decomposition of $\phi(U)$. Without loss of generality, we can choose $U^* = Q [\text{diag}(\sqrt{\boldsymbol{\lambda}}) \ \mathbf{0}]$. Then $U = Q [\text{diag}(\sqrt{\boldsymbol{\lambda}} \odot \mathbf{s}) \ \mathbf{0}] V^T$ for some orthonormal matrix $V \in \mathbb{R}^{r \times r}$ and $\mathbf{s} = [s_1 \ \cdots \ s_{r^*}]$. By Lemma 1, we get $R = V^T$. Plugging these into $DU^T = U^* R U^T - U U^T$ gives $DU^T = \mathbf{0}$.

Hence $[\nabla^2 g(U)](D, D)$ is simply determined by its first term

$$\begin{aligned} [\nabla^2 g(U)](D, D) &\leq -2\|UU^T - U^*U^{*T}\|_F^2 \\ &\leq -2(\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2 \|D\|_F^2, \end{aligned}$$

where the second line follows from Lemma 2 with $r' := \text{rank}(U)$. This further implies

$$\lambda_{\min}(\nabla^2 g(U)) \leq -2(\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2.$$

This simple example is ideal in several ways, particularly the gradient $\nabla f(\phi(U)) = \phi(U) - \phi(U^*)$, which directly establishes the negativity of the first term in (3.3); and by choosing $D = U - U^*R$ and using $DU^T = 0$, the second term vanishes. Both are not true any more for general objective functions $f(X)$. However, the example does suggest that the direction $D = U - U^*R$ is a good choice to show $[\nabla^2 g(U)](D, D) \leq -\tau \|D\|_F^2$ for some $\tau > 0$, which we will continue to use in Section 4 to prove Theorem 1.

4 Proof of Theorem 1

Proof Outline. We present a formal proof of Theorem 1 in this section. The main arguments involve showing each critical point U of $g(U)$ either corresponds to the optimal solution X^* or is a strict saddle point of $g(U)$. Being a strict saddle of $g(U)$ means the Hessian matrix $\nabla^2 g(U)$ has at least one strictly negative eigenvalue. Inspired by the discussions in Section 3, we will use the direction $D = U - U^*R$ and show that the Hessian $\nabla^2 g(U)$ has a strictly negative curvature along D : i.e., $[\nabla^2 g(U)](D, D) \leq -\tau \|D\|_F^2$, for some $\tau > 0$.

4.1 Supporting Lemmas

We first list two lemmas. Lemma 3 separates $\|(U - \hat{U})U^T\|_F^2$ into two parts: $\|UU^T - \hat{U}\hat{U}^T\|_F^2$ and $\|(UU^T - \hat{U}\hat{U}^T)QQ^T\|_F^2$ with QQ^T being the projection matrix onto $\text{Range}(U)$. It is crucial for the first part $\|UU^T - \hat{U}\hat{U}^T\|_F^2$ to have a small coefficient. In lemma 4, we will further control the second part as a consequence of U being a critical point. The proof of Lemma 3 is given in Section B.

Lemma 3. *Let U and \hat{U} be any two matrices in $\mathbb{R}^{n \times r}$ such that $U^T \hat{U} = \hat{U}^T U$ is PSD. Assume Q is an orthogonal matrix whose columns span $\text{Range}(U)$. Then*

$$\|(U - \hat{U})U^T\|_F^2 \leq \frac{1}{8}\|UU^T - \hat{U}\hat{U}^T\|_F^2 + \left(3 + \frac{1}{2(\sqrt{2} - 1)}\right)\|(UU^T - \hat{U}\hat{U}^T)QQ^T\|_F^2.$$

We remark that Lemma 3 is a strengthened version of [12, Lemma 4.4]. While the result there requires: (i) U to be a critical point of the factored objective function $g(U)$; (ii) \hat{U} is a optimal factor in \mathcal{A}^* that is closest to U , i.e., $\hat{U} = U^*R$ with $U^* \in \mathcal{A}^*$ and $R = \text{argmin}_{\tilde{R} \in \mathbb{O}_r} \|U - U^*\tilde{R}\|_F$. Lemma 3 removes these assumptions and requires only $U^T \hat{U} = \hat{U}^T U$ being PSD.

Next, we control the distance between UU^T and the global solution X^* of (1.1) when U is a critical point of the factored objective function $g(U)$, i.e., $\nabla g(U) = \mathbf{0}$. The proof, given in Section C, relies on writing $\nabla f(X) = \nabla f(X^*) + \int_0^1 \nabla^2 f(tX + (1-t)X^*)[X - X^*]dt$ and applying the R P of the Hessian matrix.

Lemma 4 (Upper bound on $\|(UU^T - X^*)QQ^T\|_F^2$). *Suppose the objective function $f(X)$ in (1.1) is twice continuously differentiable and satisfies $2r$ -restricted m -strongly convexity and M -smoothness condition (1.3). Further, let U be any critical point of (1.2) and Q be the orthonormal basis spanning $\text{Range}(U)$. Then*

$$\|(UU^T - X^*)QQ^T\|_F \leq \frac{M - m}{M + m}\|UU^T - X^*\|_F.$$

4.2 The Formal Proof

Now, we are ready to prove the main theorem.

Proof of Theorem 1. By Section 3, it suffices to find a direction D to produce a strictly negative curvature for each critical point U not corresponding to X^* . We choose $D = U - U^*R$ where $R = \text{argmin}_{\tilde{R} \in \mathbb{O}_r} \|U - U^*\tilde{R}\|_F$, $X = \phi(U)$ and $X^* = \phi(U^*)$. Then according to (3.3), we have

$$\begin{aligned} [\nabla^2 g(U)](D, D) &= 2\langle \nabla f(\phi(U)), DD^T \rangle + 4[\nabla^2 f(\phi(U))](DU^T, DU^T) \\ &\stackrel{(i)}{=} 2\langle \nabla f(X), X^* - X \rangle + 4[\nabla^2 f(X)](DU^T, DU^T) \\ &= -2 \underbrace{\langle \nabla f(X^*) - \nabla f(X), X^* - X \rangle}_{\Pi_1} + \underbrace{\langle \nabla f(X^*), X^* - X \rangle}_{\Pi_2} + 4 \underbrace{[\nabla^2 f(X)](DU^T, DU^T)}_{\Pi_3}. \end{aligned} \quad (4.1)$$

(i), as in (3.4), follows from $\nabla g(U) = 2\nabla f(UU^T)U = \mathbf{0}$ since U is a critical point of $g(U)$. Next, we bound Π_1 , Π_2 , Π_3 separately.

Bound Π_1 .

$$\begin{aligned} \Pi_1 &= \langle \nabla f(X^*) - \nabla f(X), X^* - X \rangle \stackrel{(i)}{=} \left\langle \int_0^1 [\nabla^2 f(tX + (1-t)X^*)](X^* - X)dt, X^* - X \right\rangle \\ &= \int_0^1 [\nabla^2 f(tX + (1-t)X^*)](X^* - X, X^* - X)dt \\ &\stackrel{(ii)}{\geq} m\|X^* - X\|_F^2. \end{aligned}$$

- (i) follows from the integral form of the mean value theorem for vector-valued functions (see [49, Eq. (A.57)]);
- (ii) follows from the restricted strong convexity assumption (1.3) since the PSD matrix $tX + (1-t)X^*$ has rank at most $2r$.

Bound Π_2 .

$$\Pi_2 = \langle \nabla f(X^*), X^* - X \rangle \leq 0$$

follows from the optimality condition for the convex optimization (1.1) (see, e.g., [18, Section 4.2.3]) and the fact that X^* is optimal while $X \succeq 0$ is feasible.

Bound Π_3 .

$$\Pi_3 = [\nabla^2 f(X)](DU^T, DU^T) \leq M \|DU^T\|_F^2$$

following from the restricted smoothness assumption (1.3) since $\text{rank}(X) \leq r \leq 2r$ and $X \succeq 0$. Recognizing $(U^*R)^T U = U^T U^* R \succeq 0$ by Lemma 1, we invoke Lemma 3 to bound $\|DU^T\|_F^2$ as

$$\|DU^T\|_F^2 \leq \frac{1}{8} \|X - X^*\|_F^2 + \left(3 + 1/(2(\sqrt{2} - 1))\right) \|(X - X^*)QQ^T\|_F^2.$$

Plugging Π_1, Π_2, Π_3 to (4.1), we obtain that

$$\begin{aligned} [\nabla^2 g(U)](D, D) &\leq -2m \|X^* - X\|_F^2 + 4M \left((1/8) \|X - X^*\|_F^2 + \left(3 + 1/(2(\sqrt{2} - 1))\right) \|(X - X^*)QQ^T\|_F^2 \right) \\ &\stackrel{(i)}{\leq} \left(-2m + 0.5M + 4 \left(3 + 1/(2(\sqrt{2} - 1))\right) M(M - m)/(M + m) \right) \|X^* - X\|_F^2 \\ &\stackrel{(ii)}{\leq} -0.074m \|X^* - X\|_F^2 \\ &\stackrel{(iii)}{\leq} -0.074m (\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2 \|D\|_F^2, \end{aligned}$$

where (i) follows from Lemma 4; (ii) holds for $\frac{M}{m} \leq 1.15$; (iii) follows from Lemma 2 with $r' := \text{rank}(U)$. As a consequence, we obtain

$$\lambda_{\min}(\nabla^2 g(U)) \leq \frac{[\nabla^2 g(U)](D, D)}{\|D\|_F^2} < -0.074m (\sigma_{r' \vee r^*}(U) + \sigma_{r' \vee r^*}(U^*))^2.$$

□

5 Prior Art and Inspirations

The past few years have seen a surge of interest in nonconvex reformulations of convex optimizations for efficiency and scalability reasons. Several convex optimizations of practical importance in machine learning, signal processing, and statistical problems, can be naturally formulated into nonconvex forms [39, 12, 11, 20, 37, 61, 44, 58]. Compared with the corresponding convex formulations, the nonconvex forms typically involve much fewer variables, enabling simple algorithms (*e.g.*, gradient descent [41, 31], trust-region method [60, 58], alternating optimization [6, 15]) to scale to large-scale applications.

Although these nonconvex reformulations have been known to work surprisingly well in practice, it remains an active research area to fully understand the theoretical underpinning of this phenomenon, particularly the geometrical structures of these nonconvex optimizations. The objective functions of convex optimizations have simple landscapes so that local minimizers are always global ones. However, the landscapes of general nonconvex functions can become as complicated as it could be. Even certifying the local optimality of a point for general functions is an NP-hard problem [46]. The existence of spurious local minima that are not global optima is a common issue [56, 30]. In addition, degenerate saddle points or those

surrounded by plateaus of small curvature also prevent local search algorithms from converging quickly to local optima [27].

Fortunately, for a range of convex optimizations, particularly those involving low-rank matrices, the corresponding nonconvex reformulations have nice geometric structures that allow local search algorithms to converge to global optimality [59]. Examples include low-rank matrix factorization, completion and sensing [39, 12, 11, 61, 67], tensor decomposition and completion [31, 5, 6, 38], structured element pursuit [51, 36], dictionary learning [7, 9, 8, 10, 58, 3], blind deconvolution [43, 42], phase retrieval [60, 20, 26], and many more. Based on whether smart initializations are needed, these previous works can be roughly classified into two categories. In one case, the algorithms require a problem-dependent initialization plus local refinement. A good initialization can lead to global convergence if the initial iterate lies in the attraction basins of the global optimal [39, 11, 61, 5, 6, 38, 9, 8, 10, 3, 20, 26]. For low-rank matrix recovery problems, such initializations can be obtained using spectral methods [11, 39, 11, 61, 67]; for other problems, it is more difficult to find an initial point located in the attraction basin [7, 9, 5]. The second category of works attempt to understand the empirical success of simple algorithms such as gradient descent, which converge to global optimality even with random initialization [41, 12, 31, 60, 58]. This is achieved by analyzing the objective function’s landscape and showing that they have no spurious local minima and no degenerate saddle point. Most of the works in the second category are for specific matrix sensing problems with quadratic objective functions. Our work expands this line of geometry-based convergence analysis by considering low-rank matrix optimization problems with general objective functions.

This research draws inspirations from several previous works. In [11], the authors also considered low-rank matrix optimizations with general objective functions. They characterized the local landscape around the global optima, and hence their algorithms require good initializations for global convergence. We instead characterize the global landscape by categorizing all critical points into global optima and strict saddle points. This guarantees that several local search algorithms with random initialization will converge to the global optima. Another closely related work is low-rank, PSD matrix recovery from linear observations by minimizing the factored quadratic objective function [12]. As we discussed in Section 3, low-rank and PSD matrix recovery from linear measurements is a special cases of our general objective function framework. Furthermore, by relating the first order optimality condition of the factored problem with the global optimality of the original convex program, our work provides a more transparent relationship between geometries of these two problems and greatly simplifies the theoretical argument. More recently, the authors of [16] showed that for general SDPs with linear objective functions and linear constraints, the factored problems have no spurious local minimizers. However, they did not characterize the saddle points and also did not allow nonlinear objective functions.

6 Conclusion

This work investigates the minimization of a convex function $f(X)$ over the cone of PSD matrices. To improve computational efficiency, we focus on a natural factored formulation of the original convex problem which explicitly encodes the PSD constraint. We prove that the factored problem, in spite of being nonconvex, has the following benign landscape: each critical point is either a factor of the global optimal solution to the original convex program, or a strict saddle where the Hessian matrix has a strictly negative eigenvalue. The geometric characterization of the factored objective function guarantees that many local search algorithms applied to the factored objective function converge to a global minimizer with random initializations.

References

- [1] Scott Aaronson. The learnability of quantum states. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 463, pages 3089–3114. The Royal Society, 2007.

- [2] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.
- [3] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *COLT*, pages 123–137, 2014.
- [4] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.
- [5] Anima Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models. *CoRR abs/1411.1488*, 17, 2014.
- [6] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.
- [7] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, 2014.
- [8] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.
- [9] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *COLT*, pages 779–806, 2014.
- [10] Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3858–3865, 2014.
- [11] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint*, 2015.
- [12] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.
- [13] Pratik Biswas, Tzu-Chen Liang, Kim-Chuan Toh, Yinyu Ye, and Ta-Chung Wang. Semidefinite programming approaches for sensor network localization with noisy distance measurements. *IEEE transactions on automation science and engineering*, 3(4):360, 2006.
- [14] Pratik Biswas and Yinyu Ye. Semidefinite programming for ad hoc wireless sensor network localization. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 46–54. ACM, 2004.
- [15] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.
- [16] Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *arXiv preprint arXiv:1606.04970*, 2016.
- [17] Thierry Bouwmans, Necdet Serhat Aybat, and El-hadi Zahzah. *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*. CRC Press, 2016.
- [18] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [19] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

- [20] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.
- [21] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- [22] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.
- [23] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
- [24] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
- [25] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.
- [26] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.
- [27] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.
- [28] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.
- [29] Dennis DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *Proceedings of the 23rd international conference on Machine learning*, pages 249–256. ACM, 2006.
- [30] Sean R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.
- [31] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.
- [32] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.
- [33] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.
- [34] Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3386–3393. IEEE, 2012.
- [35] Nick Higham and Pythagoras Papadimitriou. Matrix procrustes problems. *Rapport technique, University of Manchester*, 1995.
- [36] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Speeding up sum-of-squares for tensor decomposition and planted sparse vectors. *arXiv preprint arXiv:1512.02337*, 2015.
- [37] Prateek Jain, Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Computing matrix squareroot via non convex local search. *arXiv preprint arXiv:1507.05854*, 2015.

- [38] Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.
- [39] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *arXiv preprint arXiv:1605.08370*, 2016.
- [40] László Kozma, Alexander Ilin, and Tapani Raiko. Binary principal component analysis in the netflix collaborative filtering task. In *2009 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2009.
- [41] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.
- [42] Kiryung Lee and Marius Junge. Rip-like properties in subsampled blind deconvolution. *arXiv preprint arXiv:1511.06146*, 2015.
- [43] Kiryung Lee, Yihong Wu, and Yoram Bresler. Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization. *arXiv preprint arXiv:1312.0525*, 2013.
- [44] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. Overcomplete tensor decomposition via convex optimization. *arXiv preprint arXiv:1602.08614*, 2016.
- [45] Karthik Mohan and Maryam Fazel. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the 2010 American Control Conference*, pages 2953–2959. IEEE, 2010.
- [46] Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.
- [47] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.
- [48] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.
- [49] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- [50] Henrik Ohlsson, Allen Yang, Roy Dong, and Shankar Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. In *16th IFAC Symposium on System Identification, Brussels, Belgium, 11-13 July, 2012*, pages 89–94, 2012.
- [51] Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*, pages 3401–3409, 2014.
- [52] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- [53] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.
- [54] Joseph Salmon, Zachary Harmany, Charles-Alban Deledalle, and Rebecca Willett. Poisson noise reduction with non-local pca. *Journal of mathematical imaging and vision*, 48(2):279–294, 2014.
- [55] Sujay Sanghavi, Rachel Ward, and Chris D White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, pages 1–40, 2016.
- [56] Eduardo D Sontag and Héctor J Sussmann. Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, 3(1):91–106, 1989.

- [57] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2004.
- [58] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *arXiv preprint arXiv:1511.04777*, 2015.
- [59] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.
- [60] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.
- [61] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 270–289. IEEE, 2015.
- [62] Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.
- [63] M. Udell, C. Horn, S. Boyd, and R. Zadeh. Generalized low rank models. *Foundations and Trends(r) in Machine Learning*, 9(1):1–118, 2016.
- [64] Irène Waldspurger, Alexandre dAspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.
- [65] Markus Weimer, Alexandros Karatzoglou, Quoc Viet Le, and Alex Smola. Maximum margin matrix factorization for collaborative ranking. *Advances in neural information processing systems*, pages 1–8, 2007.
- [66] Qian Zhao, Deyu Meng, Zongben Xu, Wangmeng Zuo, and Lei Zhang. Robust principal component analysis with complex noise. In *Proceedings of The 31st International Conference on Machine Learning*, pages 55–63, 2014.
- [67] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.
- [68] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

A Proof of Lemma 2

Proof. Denote $X_1 = U_1 U_1^T$ and $X_2 = U_2 U_2^T$. Let X_1 and X_2 have full eigenvalue decompositions:

$$X_1 = \sum_{j=1}^n \lambda_j \mathbf{p}_j \mathbf{p}_j^T, X_2 = \sum_{j=1}^n \eta_j \mathbf{q}_j \mathbf{q}_j^T;$$

where $\{\lambda_j\}$ and $\{\eta_j\}$ are eigenvalues arranged in decreasing order. Since $\text{rank}(U_1) = r_1$ and $\text{rank}(U_2) = r_2$, we have $\lambda_j = 0$ for $j > r_1$ and $\eta_j = 0$ for $j > r_2$. We compute $\|X_1 - X_2\|_F^2$ as follows

$$\begin{aligned} \|X_1 - X_2\|_F^2 &= \|X_1\|_F^2 + \|X_2\|_F^2 - 2\langle X_1, X_2 \rangle \\ &= \sum_{i=1}^n \lambda_i^2 + \sum_{j=1}^n \eta_j^2 - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \eta_j \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 \\ &\stackrel{(i)}{=} \sum_{i=1}^n \lambda_i^2 \sum_{j=1}^n \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 + \sum_{j=1}^n \eta_j^2 \sum_{i=1}^n \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \eta_j \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 \\ &\stackrel{(ii)}{=} \sum_{i=1}^{r_1 \vee r_2} \sum_{j=1}^{r_1 \vee r_2} (\lambda_i - \eta_j)^2 \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 \\ &= \sum_{i=1}^{r_1 \vee r_2} \sum_{j=1}^{r_1 \vee r_2} (\sqrt{\lambda_i} - \sqrt{\eta_j})^2 (\sqrt{\lambda_i} + \sqrt{\eta_j})^2 \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 \\ &\stackrel{(iii)}{\geq} (\sqrt{\lambda_{r_1 \vee r_2}} + \sqrt{\eta_{r_1 \vee r_2}})^2 \sum_{i=1}^{r_1 \vee r_2} \sum_{j=1}^{r_1 \vee r_2} (\sqrt{\lambda_i} - \sqrt{\eta_j})^2 \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 \\ &\stackrel{(iv)}{=} (\sqrt{\lambda_{r_1 \vee r_2}} + \sqrt{\eta_{r_1 \vee r_2}})^2 \|\sqrt{X_1} - \sqrt{X_2}\|_F^2, \end{aligned}$$

where (i) follows from $\sum_{j=1}^n \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 = \|\mathbf{p}_i\|_2^2 = 1$ since $\{\mathbf{q}_j\}$ form an orthonormal basis and similarly $\sum_{i=1}^n \langle \mathbf{p}_i, \mathbf{q}_j \rangle^2 = \|\mathbf{q}_j\|_2^2 = 1$; (ii) follows from exchanging the summations, the fact that $\lambda_j = 0$ for $j > r_1$ and $\eta_j = 0$ for $j > r_2$, and completing squares; (iii) follows from $\{\lambda_j\}$ and $\{\eta_j\}$ are sorted in decreasing order. (iv) follows from observing (ii) and $\{\sqrt{\lambda_j}\}$ and $\{\sqrt{\eta_j}\}$ are eigenvalues of $\sqrt{X_1}$ and $\sqrt{X_2}$, the matrix square root of X_1 and X_2 , respectively.

Finally, we can conclude the proof as long as we can show the following inequality:

$$\|\sqrt{X_1} - \sqrt{X_2}\|_F^2 \geq \min_{R \in \mathbb{O}_r} \|U - \hat{U}R\|_F^2. \quad (\text{A.1})$$

By expanding $\|\cdot\|_F^2$ using inner products, and noting $\langle \sqrt{X_1}, \sqrt{X_1} \rangle = \text{trace}(X_1) = \text{trace}(U_1 U_1^T)$ and $\langle \sqrt{X_2}, \sqrt{X_2} \rangle = \text{trace}(X_2) = \text{trace}(U_2 U_2^T)$, (A.1) reduces to

$$\langle \sqrt{X_1}, \sqrt{X_2} \rangle \leq \max_{R \in \mathbb{O}_r} \langle U_1, U_2 R \rangle. \quad (\text{A.2})$$

To show (A.2), we write the SVD of U_1, U_2 as $U_1 = P_1 \Sigma_1 Q_1^T$ and $U_2 = P_2 \Sigma_2 Q_2^T$ with $P_1, P_2 \in \mathbb{R}^{n \times r}$, $\Sigma_1, \Sigma_2 \in \mathbb{R}^{r \times r}$ and $Q_1, Q_2 \in \mathbb{R}^{r \times r}$. Then we have $\sqrt{X_1} = P_1 \Sigma_1 P_1^T$, $\sqrt{X_2} = P_2 \Sigma_2 P_2^T$. On one hand, we can compute the RHS of (A.2):

$$\begin{aligned} \text{RHS} &= \max_{R \in \mathbb{O}_r} \langle P_1 \Sigma_1 Q_1^T, P_2 \Sigma_2 Q_2^T R \rangle \\ &= \max_{R \in \mathbb{O}_r} \langle P_1 \Sigma_1, P_2 \Sigma_2 Q_2^T R Q_1 \rangle \\ &\stackrel{(i)}{=} \max_{R \in \mathbb{O}_r} \langle P_1 \Sigma_1, P_2 \Sigma_2 R \rangle \\ &\stackrel{(ii)}{=} \|(P_2 \Sigma_2)^T P_1 \Sigma_1\|_*, \end{aligned}$$

where (i) follows by a change of variable $R = Q_2^T R Q_1$; (ii) follows from Lemma 1. On the other hand, the LHS of (A.2) simplifies as

$$\begin{aligned} \text{LHS} &= \langle P_1 \Sigma_1 P_1^T, P_2 \Sigma_2 P_2^T \rangle \\ &= \langle (P_2 \Sigma_2)^T P_1 \Sigma_1, P_2^T P_1 \rangle \\ &\stackrel{(i)}{\leq} \|(P_2 \Sigma_2)^T P_1 \Sigma_1\|_* \|P_2^T P_1\| \\ &\stackrel{(ii)}{\leq} \|(P_2 \Sigma_2)^T P_1 \Sigma_1\|_*, \end{aligned}$$

where (i) follows from Hölder's inequality and (ii) follows since $\|P_2^T P_1\| \leq \|P_2\| \|P_1\| \leq 1$. This proves (A.2). Finally, we conclude the proof by noting that $\sqrt{\lambda_{r_1 \vee r_2}} = \sigma_{r_1 \vee r_2}(U_1)$ and $\sqrt{\eta_{r_1 \vee r_2}} = \sigma_{r_1 \vee r_2}(U_2)$, where $\sigma_i(\cdot)$ denotes the i th largest singular value of its argument. \square

B Proof of Lemma 3

The proof for Lemma 3 is inspired by that in [12] and one needs the following lemma.

Lemma 5. [12, Lemma E.1] *Let U and \hat{U} be any two matrices in $\mathbb{R}^{n \times r}$ such that $U^T \hat{U} = \hat{U}^T U$ is PSD. Then*

$$\left\| (U - \hat{U}) U^T \right\|_F^2 \leq \frac{1}{2(\sqrt{2} - 1)} \left\| U U^T - \hat{U} \hat{U}^T \right\|_F^2.$$

Proof of Lemma 3. Define two orthogonal projections $\mathbb{Q} = Q Q^T$ that projects onto the range space of U and $\mathbb{Q}_\perp = Q_\perp Q_\perp^T$ that projects onto its orthogonal complement. Denote $\tau = 1/(2\sqrt{2} - 2)$. Then

$$\begin{aligned} &\|(U - \hat{U}) U^T\|_F^2 \\ &= \|(U - \mathbb{Q} \hat{U}) U^T\|_F^2 + \|\mathbb{Q}_\perp \hat{U} U^T\|_F^2 \\ &\stackrel{(i)}{\leq} \tau \|U U^T - \mathbb{Q} \hat{U} \hat{U}^T\|_F^2 + \langle \hat{U}^T \mathbb{Q}_\perp \hat{U}, U^T U - \hat{U}^T \mathbb{Q} \hat{U} \rangle + \langle \hat{U}^T \mathbb{Q}_\perp \hat{U}, \hat{U}^T \mathbb{Q} \hat{U} \rangle \\ &\stackrel{(ii)}{\leq} \tau \|U U^T - \mathbb{Q} \hat{U} \hat{U}^T\|_F^2 + (1/8) \|\hat{U}^T \mathbb{Q}_\perp \hat{U}\|_F^2 + 2 \|U^T U - \hat{U}^T \mathbb{Q} \hat{U}\|_F^2 + \langle \hat{U}^T \mathbb{Q}_\perp \hat{U}, \hat{U}^T \mathbb{Q} \hat{U} \rangle, \end{aligned} \tag{B.1}$$

where (i) follows from Lemma 5 by noticing $U^T \mathbb{Q} \hat{U} = U^T \hat{U} \succeq 0$ and $\|\mathbb{Q}_\perp \hat{U} U^T\|_F^2 = \langle \hat{U}^T \mathbb{Q}_\perp \hat{U}, U^T U \rangle$; (ii) follows from $ab \leq a^2/8 + 2b^2$.

Show $\|\hat{U}^T \mathbb{Q}_\perp \hat{U}\|_F^2 \leq \|\hat{U} \hat{U}^T - U U^T\|_F^2$.

$$\|\hat{U}^T \mathbb{Q}_\perp \hat{U}\|_F^2 = \langle \hat{U} \hat{U}^T \mathbb{Q}_\perp, \mathbb{Q}_\perp \hat{U} \hat{U}^T \rangle = \|\mathbb{Q}_\perp \hat{U} \hat{U}^T \mathbb{Q}_\perp\|_F^2 \stackrel{(i)}{=} \|\mathbb{Q}_\perp (\hat{U} \hat{U}^T - U U^T) \mathbb{Q}_\perp\|_F^2 \stackrel{(ii)}{\leq} \|\hat{U} \hat{U}^T - U U^T\|_F^2.$$

(i) follows from $\mathbb{Q}_\perp U = 0$; (ii) follows from nonexpansiveness of projection operator.

Show $\langle \hat{U}^T \mathbb{Q}_\perp \hat{U}, \hat{U}^T \mathbb{Q} \hat{U} \rangle \leq \|\mathbb{Q} \hat{U} \hat{U}^T - U U^T\|_F^2$.

$$\langle \hat{U}^T \mathbb{Q}_\perp \hat{U}, \hat{U}^T \mathbb{Q} \hat{U} \rangle = \langle \mathbb{Q} \hat{U} \hat{U}^T, \hat{U} \hat{U}^T \mathbb{Q}_\perp \rangle = \|\mathbb{Q} \hat{U} \hat{U}^T \mathbb{Q}_\perp\|_F^2 \stackrel{(i)}{=} \|\mathbb{Q} (\hat{U} \hat{U}^T - U U^T) \mathbb{Q}_\perp\|_F^2 \stackrel{(ii)}{\leq} \|\mathbb{Q} \hat{U} \hat{U}^T - U U^T\|_F^2.$$

(i) follows from $\mathbb{Q}_\perp U = 0$; (ii) follows from nonexpansiveness of projection operator.

Show $\|U^T U - \hat{U}^T \mathbb{Q} \hat{U}\|_F^2 \leq \|U U^T - \mathbb{Q} \hat{U} \hat{U}^T\|_F^2$. By expanding $\|\cdot\|_F^2$ using inner products, first note that $\|U^T U\|_F^2 = \|U U^T\|_F^2$ and $\|\hat{U}^T \mathbb{Q} \hat{U}\|_F^2 = \|\mathbb{Q} \hat{U} \hat{U}^T \mathbb{Q}\|_F^2 \leq \|\hat{U} \hat{U}^T \mathbb{Q}\|_F^2$. Hence, it reduces to showing $\langle U^T U, \hat{U}^T \mathbb{Q} \hat{U} \rangle \geq \langle U U^T, \mathbb{Q} \hat{U} \hat{U}^T \rangle$, or equivalently, showing

$$\langle U^T U, \hat{U}^T \mathbb{Q} \hat{U} \rangle \geq \|\hat{U}^T U\|_F^2.$$

The caveat is using SVD of U to rewrite $U^T U$ and \mathbb{Q} . Let $\text{rank}(U) = r'$ and SVD of $U = Q\Sigma P^T$ with $\Sigma \in \mathbb{R}^{r' \times r'}$ and $P \in \mathbb{R}^{r \times r'}$. Then $U^T U = P\Sigma^2 P^T$, $Q = UP\Sigma^{-1}$ and $\mathbb{Q} = QQ^T = UP\Sigma^{-2}P^T U^T$. Then

$$\begin{aligned} \langle U^T U, \hat{U}^T \mathbb{Q} \hat{U} \rangle &= \langle P\Sigma^2 P^T, \hat{U}^T UP\Sigma^{-2}P^T U^T \hat{U} \rangle \\ &\stackrel{(i)}{=} \langle \Sigma^2, P^T (U^T \hat{U}) P \Sigma^{-2} P^T (U^T \hat{U}) P \rangle \\ &\stackrel{(ii)}{=} \langle \Sigma^2, G \Sigma^{-2} G \rangle = \|\Sigma G \Sigma^{-1}\|_F^2 \stackrel{(iii)}{\geq} \|G\|_F^2 \stackrel{(iv)}{=} \|U^T \hat{U}\|_F^2, \end{aligned}$$

where (i) follows from $\hat{U}^T U = U^T \hat{U}$; (ii) follows from $G := P^T (U^T \hat{U}) P$ and note that $G \succeq 0$ since $U^T \hat{U} \succeq 0$. (iii) follows from $\|\Sigma G \Sigma^{-1}\|_F^2 = \sum_{i,j \in [r']} \frac{\sigma_i^2}{\sigma_j^2} G_{ij}^2$, $G_{ij} = G_{ji}$ and $a^2 + b^2 \geq 2ab$. (iv) follows from $\|G\|_F^2 = \langle P P^T (U^T \hat{U}), (\hat{U}^T U) P P^T \rangle = \|U^T \hat{U}\|_F^2$ since $U P P^T = U$.

Finally, We concludes the proof by plugging these three bounds into (B.1). \square

C Proof of Lemma 4

Proof of Lemma 4. Let $X = \phi(U)$ and $X^* = \phi(U^*)$. Denote $\delta = \frac{M-m}{M+m}$. We start with the critical point condition $\nabla f(X)U = \mathbf{0}$.

$$\begin{aligned} &\Rightarrow \nabla f(X)U = \mathbf{0} \\ &\Rightarrow \nabla f(X)QQ^T = \mathbf{0} \\ &\Rightarrow \langle \nabla f(X), ZQQ^T \rangle = 0, \forall Z \\ &\stackrel{(i)}{\Rightarrow} \langle \nabla f(X^*) + \int_0^1 \nabla^2 f(tX + (1-t)X^*)[X - X^*]dt, ZQQ^T \rangle = 0, \forall Z \\ &\Rightarrow \langle \nabla f(X^*), ZQQ^T \rangle + \int_0^1 [\nabla^2 f(tX + (1-t)X^*)](X - X^*, ZQQ^T)dt = 0, \forall Z \\ &\stackrel{(ii)}{\Rightarrow} |-2/(M+m)\langle \nabla f(X^*), ZQQ^T \rangle - \langle X - X^*, ZQQ^T \rangle| \leq \delta \|X - X^*\|_F \|ZQQ^T\|_F, \forall Z \\ &\Rightarrow |2/(M+m)\langle \nabla f(X^*), ZQQ^T \rangle + \langle X - X^*, ZQQ^T \rangle| \leq \delta \|X - X^*\|_F \|ZQQ^T\|_F, \forall Z \\ &\stackrel{(iii)}{\Rightarrow} |2/(M+m)\langle \nabla f(X^*), (X - X^*)QQ^T \rangle + \|(X - X^*)QQ^T\|_F^2| \leq \delta \|X - X^*\|_F \|(X - X^*)QQ^T\|_F. \end{aligned}$$

(i) follows from the integral form of the mean value theorem for vector-valued functions (see [49, Eq. (A.57)]); (ii) from $\widehat{\text{RIP}}$ by noting that the PSD matrix $tX^* + (1-t)X$ has rank at most $2r$. (iii) follows by choosing $Z = X - X^*$.

We conclude the proof by noting that $\langle \nabla f(X^*), (X - X^*)QQ^T \rangle \geq 0$ since

$$\langle \nabla f(X^*), (X - X^*)QQ^T \rangle \stackrel{(i)}{=} \langle \nabla f(X^*), X \rangle \stackrel{(ii)}{\geq} 0.$$

(i) follows from $\nabla f(X^*)X^* = X^* \nabla f(X^*) = \mathbf{0}$ and $XQQ^T = X$; (ii) follows from $\nabla f(X^*) \succeq 0, X \succeq 0$. \square